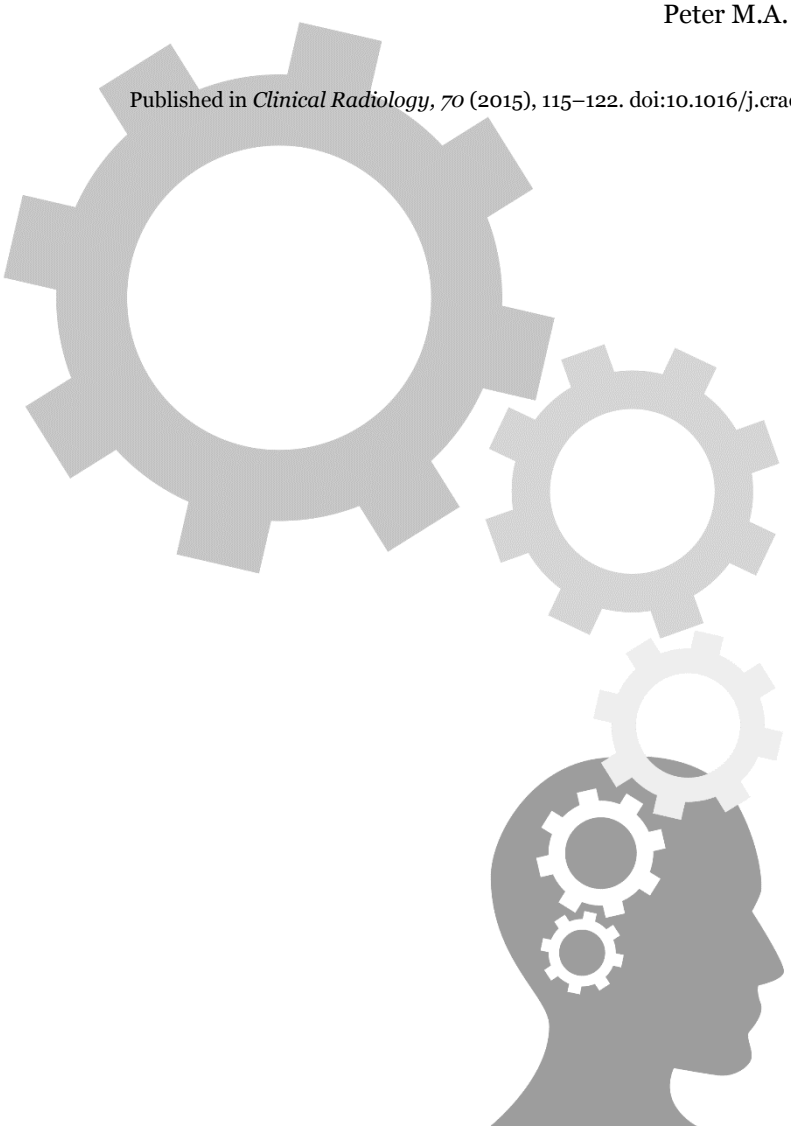
7

Improving the radiologist-CAD interaction: designing for appropriate trust

Wiard Jorritsma
Fokie Cnossen
Peter M.A. van Ooijen

Published in *Clinical Radiology, 70* (2015), 115–122. doi:10.1016/j.crad.2014.09.017



Abstract

Computer-aided diagnosis (CAD) has great potential to improve radiologists' diagnostic performance. However, the reported performance of the radiologist–CAD team is lower than what might be expected based on the performance of the radiologist and the CAD system in isolation. This indicates that the interaction between radiologists and the CAD system is not optimal. An important factor in the interaction between humans and automated aids (such as CAD) is trust. Suboptimal performance of the human–automation team is often caused by an inappropriate level of trust in the automation. In this review, we examine the role of trust in the radiologist–CAD interaction and suggest ways to improve the output of the CAD system so that it allows radiologists to calibrate their trust in the CAD system more effectively. Observer studies of the CAD systems show that radiologists often have an inappropriate level of trust in the CAD system. They sometimes under-trust CAD, thereby reducing its potential benefits, and sometimes over-trust it, leading to diagnostic errors they would not have made without CAD. Based on the literature on trust in human–automation interaction and the results of CAD observer studies, we have identified four ways to improve the output of CAD so that it allows radiologists to form a more appropriate level of trust in CAD. Designing CAD systems for appropriate trust is important and can improve the performance of the radiologist–CAD team. Future CAD research and development should acknowledge the importance of the radiologist–CAD interaction, and specifically the role of trust therein, in order to create the perfect artificial partner for the radiologist. This review focuses on the role of trust in the radiologist–CAD interaction. The aim of the review is to encourage CAD developers to design for appropriate trust and thereby improve the performance of the radiologist–CAD team.

*The key to winning the race is not to compete **against** machines but to compete **with** machines.*

— Erik Brynjolfsson & Andrew McAfee

Introduction

Medical image diagnosis is a highly complex task with very demanding cognitive and perceptual components. Radiologists have developed the ability to perform this task with impressive accuracy and efficiency. However, no matter how skilled the radiologist, he or she is never immune to making errors. Various studies have documented the occurrence of radiological errors in clinical practice, ranging from missed lesions due to perceptual oversight to the incorrect recommendation of follow-up procedures (e.g. [1–9]).

Computer-aided diagnosis

Computer-aided diagnosis (CAD), in which sophisticated image processing and artificial intelligence techniques are used to detect and/or evaluate abnormalities in medical images, has great potential to improve radiologists' diagnostic performance. CAD can be used as a second opinion: drawing radiologists' attention to abnormalities they overlooked or prompting them to reevaluate structures they initially diagnosed incorrectly.

A distinction can be made between computer-aided detection (CADe), which focuses on the *detection* of abnormalities, and computer-aided diagnosis (CADx), which focuses on the *diagnosis* of abnormalities. CADe systems identify and mark abnormal regions in a medical image. Radiologists first perform an unaided reading of the image and then review the marks made by the CADe system. Fig. 1 shows an example of a typical CADe system for lung nodule detection on chest radiographs.

CADx systems focus on diagnosis rather than detection. Suspicious structures within an image are identified by the radiologist and evaluated by the CADx system. This evaluation can be a decision whether the structure is benign or malignant, the estimated likelihood of malignancy, or a pathological classification.

Because many systems perform both detection and diagnosis, the distinction between CADe and CADx is not always clear. Differentiating between the two types of CAD is also not very relevant for the purpose of this review. We will therefore use the general term CAD for both CADe and CADx systems.

Radiologist-CAD interaction

The combination of a radiologist and a CAD system constitutes a *diagnostic team*, in the same way that two radiologists in a double reading setting do. The performance of this team is determined by the individual performance of its "members" and the quality of interaction between the members. Various studies have shown that radiologists and CAD can make an effective team that reaches a higher level of diagnostic performance than one radiologist alone (e.g. [10–15]).

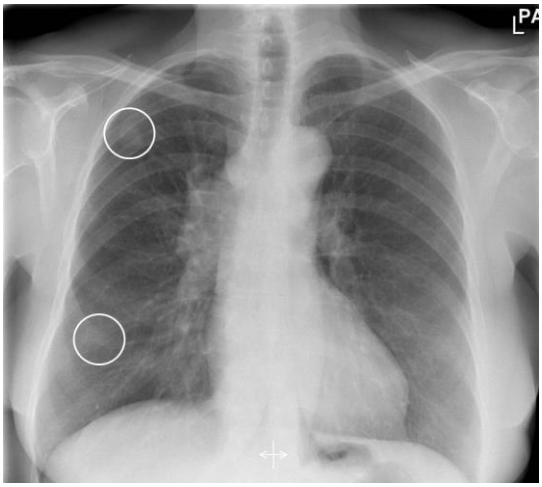


Figure 1. Example of a typical CADe system for lung nodule detection on chest radiographs. The circles are marks made by the system to indicate the presence of a lung nodule. The upper mark is a true positive. The lower mark is a false positive.

However, the team performance of radiologist and CAD is lower than what might be expected based on the performance of the radiologist and the CAD system in isolation [16,17]. There are even studies that found no benefits of CAD on radiologists' diagnostic performance (e.g. [18–23]), an increased sensitivity at the cost of reduced specificity (e.g. [24,25]), or even reduced sensitivity of the best performing radiologists for difficult cases [26]. This suggests that the interaction between radiologists and CAD is not optimal.

An important factor in the interaction between humans and automated aids (such as CAD) is trust [27–33]. The more humans trust an automated aid, the more likely they are to rely on its decisions. For optimal performance of the human-automation team, it is vital that an appropriate level of reliance occurs. However, humans often show an inappropriate level of automation reliance caused by an inappropriate amount of trust in the automation [33]. Too little trust in a useful automated aid can lead to underreliance, which means that the full potential of the aid is not being used. Too much trust in an aid on the other hand can lead to overreliance, meaning that the aid causes humans to make errors they would not have made without it [34].

Inappropriate trust in CAD

Under-trust

Too little trust in a useful automated aid can lead to disuse, i.e. an underreliance on automation [34]. Disuse has been found in a wide variety of settings. For example, Dzindolet et al. [35] and Beck et al. [36] found that participants performing a detection task in which they had to decide whether a camouflaged soldier was present in a scene often ignored the decisions of an automated aid, even when they knew the aid had superior performance on the task. Wang et al. [37] found that participants sometimes ignored the true negatives of a Combat Identification System, even though they were informed that the system did not make any false negative errors. Disuse was also found in early versions of the Ground Proximity Warning System, which pilots did not trust due to its high false alarm rate [34].

There are various examples of disuse in the CAD literature. Several studies have shown that radiologists ignore a substantial amount of true positive CAD marks (20% in [38], 22% in [20], and 33% in [39] of the total amount of true positives, and 53% in [40], 71% in [17], and 84% in [23] of the true positives they overlooked during unaided reading).

Similarly, Halligan et al. [41] found that the sensitivity of radiologists performing polyp detection on CT colonography images increased for small- and medium-sized polyps correctly marked by CAD, but not for correctly marked large polyps. This tendency to ignore correctly marked large polyps was also found by Taylor et al. [42].

Shiraishi et al. [43] found that a CAD system designed to help radiologists distinguish benign from malignant lung nodules improved their diagnostic performance, but the performance of the radiologist-CAD team was lower than the standalone performance of the CAD system, indicating that radiologists underutilized the CAD system's capabilities.

Over-trust

Too much trust in automation, on the other hand, can lead to misuse, i.e. an overreliance on automation [34]. Like disuse, misuse has been found in various domains. For example, Skitka et al. [44] found that participants performing a simulated flight task with an automated aid performed worse than unaided participants. This was caused by an overreliance on the aid, which led participants to make false negative errors when the aid missed an event and false positive errors when the aid made an incorrect recommendation (even when it contradicted their training and other available information).

Itoh [45] found an overreliance on an adaptive cruise control system in a driving simulator. Even though participants were informed that the system had a limited deceleration rate, four out of twelve participants caused a rear-end collision because they trusted the system to brake in time.

There are also examples of misuse in the CAD literature. Various studies have found that CAD decreased radiologists' specificity (e.g. [24,25]). This indicates that radiologists put too much trust in the CAD marks, causing them to accept a substantial number of CAD false positives. Lee et al. [40] found that radiologists performing lung nodule detection on chest radiographs accepted one false positive mark per 50 images and residents accepted one false positive mark per nine images. In [23], radiologists performing the same task accepted one false positive per 19 images and residents one per 11 images.

Alberdi et al. [46] found that radiologists assisted by CAD had a lower sensitivity than unaided radiologists on a mammography data set containing a large proportion of cancers missed by CAD. This indicates that radiologists put too much trust in the CAD system's ability to detect abnormalities, which led them to revise a substantial amount of their true positive decisions of certain structures based on the (incorrect) absence of CAD marks on these structures.

Povyakalo et al. [47] found that CAD improved radiologists' sensitivity for breast cancers that were relatively easy to detect, but decreased sensitivity for cancers that were relatively difficult to detect. This was probably caused by the fact that radiologists were more likely to rely on CAD for difficult cancers, because they were less certain of their own decisions. However, cancers that were difficult for radiologists were also difficult for CAD, leading to a large number of CAD false negative errors for these cancers, which made a high level of reliance on CAD for difficult cancers inappropriate. A similar result was found in another study [26], where CAD decreased the sensitivity of the best performing radiologists for difficult cases.

The potential for misuse is even greater when CAD is used as a concurrent reader (i.e. CAD output is immediately available to the radiologists) instead of as a second reader (i.e. CAD output is only available after radiologists have viewed the image on their own). Zheng et al. [21][48] found that a poorly performing CAD system used as a concurrent reader significantly decreased radiologists' diagnostic performance in mammogram reading. Beyer et al. [49] found that concurrent reader CAD decreased radiologists' sensitivity for detecting lung nodules in CT scans. These results suggest that radiologists may over-trust CAD's sensitivity. When the CAD marks are presented at the onset of image reading, radiologists focus on these marks and pay less attention to unmarked regions. This causes them to miss abnormalities that were not marked by CAD.

Designing for appropriate trust

The results of observer studies with CAD show that radiologists often have an inappropriate level of trust in CAD. We believe that the reason for this is that the output of CAD systems is often presented to radiologists in such a way that it is impossible for them to establish an optimal level of trust in the system. In this section, we will suggest ways to improve CAD's output so that it allows radiologists to calibrate their trust in CAD more effectively.

Confidence rating

Like radiologists, CAD systems have a response criterion. When the information obtained from a certain structure within an image exceeds this criterion the structure is considered abnormal. When the information does not meet the criterion it is considered normal. Most CAD systems do not differentiate between structures that exceed the response criterion by a large amount (for which CAD has a high "confidence" that they are abnormal), and structures that barely exceed the criterion (for which CAD has a low confidence that they are abnormal): the system either does or does not mark the structure.

Displaying a confidence rating for each mark might facilitate more appropriate trust, because it allows radiologists to adapt their trust in a specific mark to CAD's confidence in this mark. This could lead to less disuse, which is often associated with systems that have a high false positive rate [34], as most CAD systems do, because false positive marks likely have a smaller negative effect on trust in the entire system when radiologists know that the system did not consider it likely that the marked region was in fact abnormal. Their trust in CAD marks that have a high likelihood of abnormality could then remain at a high level, causing them to dismiss a smaller number of true positive marks. It could also lead to less misuse, because radiologists are probably less inclined to trust false positive CAD marks when they know that CAD does not have a high confidence in these marks.

There are CAD systems that do present the confidence ratings of their marks to radiologists. For example, in the system used by Gilbert et al. [50] the size of a mark corresponded to the likelihood of cancer as determined by the system. However, this study did not compare performance of radiologists with and without CAD, or the effectiveness of CAD with and without likelihood ratings.

In the CAD system used by Shiraishi et al. [51] the color of a mark represented its likelihood of malignancy. A discrete five-color scale was used, ranging from green (a low likelihood of malignancy) to red (a high likelihood of malignancy). This system improved radiologists' diagnostic performance, but the effects of the likelihood ratings were not evaluated.

The CAD system used by Taylor et al. [20] emphasized marks in which it was particularly confident. Radiologists were significantly less likely to ignore these emphasized marks compared to regular marks. This indicates that radiologists were more likely to trust a CAD mark if they knew that CAD had a high confidence in this mark. However, it is also possible that they simply accepted more emphasized marks because these marks had more obvious signs of abnormality compared to marks in which CAD was less confident.

In the CAD setup used by Samulski et al. [52], radiologists could request CAD information of specific image regions by clicking on them. If present, a CAD mark and its associated malignancy score were displayed. The malignancy score was represented by a continuous color scale, ranging from yellow to red (low to high malignancy score). This system improved radiologists' diagnostic performance compared to unaided reading. In a subsequent study, Hupse et al. [53] found that radiologists performed better with this interactive CAD system than with a conventional CAD system that only displayed marks. However, the fact that one CAD system was interactive and the other was not prevents us from concluding that the increase in performance was due to the malignancy ratings.

Several studies [54–58] found that radiologists' classification performance of breast masses improved when they were assisted by a CAD system that provided a malignancy rating of the masses. Similar results have been found for lung nodule classification [13,59,60].

These results show that CAD confidence ratings can be useful, but there are no studies that compare the effectiveness of a CAD system that provides a confidence rating for its decisions and a CAD system that only provides discrete decisions. However, the effectiveness of displaying automation confidence has been demonstrated in other domains. For example, pilots performing a navigation and collision avoidance task [61], an anti-aircraft battle task [61], and a flight task under icing conditions [62] have been shown to form a more appropriate level of trust in an automated aid when it displayed the degree of certainty in its decisions compared to when this information was not displayed. This improvement in trust calibration led to an increase in performance of the human-automation team. Similar results were found when displaying uncertainty information in an adaptive cruise control system [63].

Rationale

Because humans and computers make decisions in a different way, it is sometimes difficult for a human to understand why an automated aid has made a certain decision. Automation errors that seem obvious to a human observer can have a negative impact on trust in the aid and are often used as justification for disuse [33,64]. In an ethnographic study of CAD usage [65], radiologists indicated that

CAD often marked the *"wrong things"* – benign features and artifacts of the image production process – and often missed obvious lesions. This led to distrust and a lack of understanding of the CAD system.

To facilitate appropriate trust in CAD, it is therefore important that radiologists have a sufficient understanding of CAD's decision making process. This can be realized in two ways: (1) by informing radiologists of the global workings of CAD's algorithms and the limitations of these algorithms (a *global rationale*), and (2) by providing an explanation for each specific CAD decision (a *local rationale*).

Global rationale

The global rationale approach requires the development of adequate instructions for CAD. The goal of the instructions is to inform radiologists of the mechanisms that determine CAD's behavior and the specific circumstances in which it is likely to make an error. These instructions should be part of the CAD training radiologists receive.

This approach has the potential to increase trust in CAD, because the negative impact of obvious CAD errors on trust might be reduced when radiologists understand the cause of these errors. For example, some CAD systems place an upper bound on the size of structures they consider for evaluation [66]. When radiologists are aware of this, the observation that CAD systematically misses extremely obvious large lesions likely has a smaller negative impact on their trust in CAD than when they are not aware of this limitation.

In addition to its potential to reduce disuse by mitigating the negative effects on trust of obvious errors, an understanding of CAD's algorithms can also reduce misuse. For example, radiologists who correctly detect a nodule near the chest wall but are not very confident in their decision might be less inclined to change their decision based on a false negative CAD decision for this structure when they know that CAD is likely to miss nodules near the chest wall.

Although the global rationale approach seems intuitive, there is (to the best of our knowledge) only one study that has empirically evaluated its effectiveness. This study found that participants provided with a brief explanation of the workings of an automated soldier detection aid and the circumstances in which it was likely to make an error, trusted the aid more than participants who were not provided with this explanation [33]. However, this effect was independent of the aid's performance level, so the rationale caused a more appropriate level of trust for participants paired with a high performance aid, but over-trust for participants paired with a low performance aid. This indicates that great care should be taken in formulating the rationale, so that it facilitates appropriate trust and not just more trust.

Bahner et al. [67] found that participants provided with a textual explanation of the circumstances in which an automated aid was likely to err during training trusted the aid more than participants provided with examples of the aid's errors. This shows that the means of presenting the rationale influences trust and underlines the importance of formulating the rationale in an appropriate way.

Local rationale

Even when radiologists have a global understanding of CAD's algorithms and their behavior, the rationale for each specific CAD decision is still formed by speculation (albeit more informed speculation) of why CAD made this decision. Another (possibly complementary) approach to facilitating a better understanding of CAD's decisions is to provide a rationale for each specific decision. This local rationale approach has been implemented in automated aids from various domains (e.g. medical diagnosis [68], financial auditing [69,70] and entertainment recommendation [71–73]). Users generally prefer seeing the local rationales compared to no rationales and providing these rationales has been shown to increase objective [70,72] and subjective measures of trust [69,73].

These results suggest that the local rationale approach might also be useful for CAD. However, because CAD arrives at a decision in a highly complex way that can be very different from the way a radiologist reaches a decision, it is difficult to display CAD's rationale for a specific decision in a way that makes sense to a radiologist.

A typical CAD system bases its decisions on a set of features it extracts from the image. For each decision, the subset of features that contributed most to the decision could be presented to the radiologist. In this way, the radiologist can truly assess why CAD diagnosed a certain structure as abnormal. However, not all features that are relevant for CAD might be meaningful to the radiologist.

An example of a system that provides such a type of local rationale is PeerView Digital: an extended feature of Hologic's ImageChecker CAD for mammography [74]. It enhances the marked regions to help radiologists visualize and analyze its specific features. It outlines the central density of detected masses and distortions and highlights detected microcalcifications. An example of its output is shown in Fig. 2.

Performance level

Because appropriate trust occurs when the level of trust matches CAD's performance level, explicitly informing radiologists of CAD's past performance seems like a natural way to improve trust calibration. Because CAD's performance can differ greatly between different types of lesions, the performance level should be presented in a way that differentiates between lesion types. This could allow

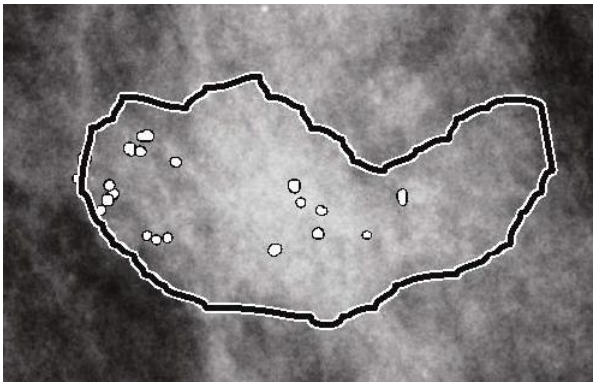


Figure 2. Output of the PeerView Digital feature of Hologic's ImageChecker CAD. This example shows the enhancement of a "Malc" CAD mark (a region containing features of both mass and calcification). The system outlines the mass and highlights the individual microcalcifications.

radiologists to calibrate their trust for each specific type of lesion and might reduce both disuse and misuse by mitigating the effects of negative and positive CAD experiences for one lesion type on trust in CAD for other lesion types.

CAD's performance can also vary depending on the image acquisition protocol that was used to scan the patient. For example, the performance of the CAD system used by [15] was greatly influenced by the slice thickness of the CT scan, with sensitivity dropping from 81% for 0.9-mm slices to 51% for 3-mm slices. Similar findings were reported in [75,76]. When working with a CAD system that has such a large variability in performance, it is vital that radiologists are informed of the performance differences between different contexts of operation.

Consider for example a radiologist that uses CAD for lung nodule detection on CT. This radiologist usually reviews thin-slices images and his trust in CAD is calibrated at a level that is appropriate for these types of images. Occasionally he reviews images with thicker slices. Now if he is unaware of the effect of slice thickness on CAD performance, his trust in CAD when reviewing thick-slice images will be at a level that is appropriate for thin-slice images but is highly inappropriate in the current situation.

To ensure that the performance information radiologists receive is accurate, it is important that the CAD system is evaluated using a valid data set and a valid gold standard against which its output is compared. While this sounds obvious, conducting a CAD performance study is far from trivial and there are currently no standardized approaches for evaluating and reporting CAD performance levels [77].

To the best of our knowledge, there is only one study that has evaluated the effects of informing users of automation performance on trust in the automation. This study showed that providing participants with a performance measure of an

automated aid for camouflaged soldier detection facilitated more appropriate trust in the aid [33]. However, this effect was only found when participants were not allowed to view the aid's decisions (they had to trust the aid blindly), which is, as the authors also note, an unrealistic scenario.

Discussion

In this review, we have shown that radiologists often have an inappropriate level of trust in CAD, which leads to suboptimal performance of the radiologist-CAD team. Radiologists sometimes under-trust CAD, thereby reducing its potential benefits, and sometimes over-trust it, leading to diagnostic errors they would not have made without CAD. We have identified four ways to improve CAD's output so that it allows radiologists to form a more appropriate level of trust in CAD: (1) presenting a confidence rating for CAD's decisions, (2) providing a global rationale for CAD's decision making process, (3) providing a local rationale for each specific CAD decision, and (4) informing radiologists of CAD's performance levels in different contexts.

There is substantial empirical evidence indicating that providing radiologists with these sources of information can facilitate more appropriate trust in CAD and thereby improve the performance of the radiologist-CAD team. However, all evidence to date is circumstantial. More research is needed to determine whether the suggested changes truly improve trust calibration and to determine the most effective way of presenting the information to the radiologists.

A realistic future possibility is that radiologists work with multiple CAD systems for different diagnostic tasks. This makes it even more important to design for appropriate trust, because radiologists need to be able to adjust their trust to the different systems. In this situation it is especially important that radiologists understand the rationale behind each CAD system's decisions and their individual performance levels so that they are aware of each system's specific strengths and limitations and do not overgeneralize positive or negative experiences from one system to the others.

Most research on CAD focuses on improving CAD's performance. While this is obviously important, the power of this research is not fully harnessed if the increase in performance is not matched by an equivalent increase in trust. Simply making CAD more trustworthy does not guarantee that it is actually trusted more. Research on the radiologist-CAD interaction, and specifically the role of trust therein, is therefore also of paramount importance. Without more research in this area, the performance of the radiologist-CAD team will never reach its maximum level.

When designing for appropriate trust, usability principles also need to be taken into consideration. Design choices that improve trust calibration should not interfere too much with the CAD system's usability and the clinical workflow of the radiologist. Poor usability is likely to result in less use of the system.

It is also possible that usability influences trust; if a CAD system is tedious to use, radiologists might be less likely to trust it. [78] found a strong correlation between a book vendor's website usability and participants' trust in the vendor. Although most components of trust measured in this study are irrelevant for the radiologist-CAD interaction, the one component that was relevant, the perceived ability of the vendor, was most strongly influenced by the website's usability.

Two of our suggestions for more appropriate trust calibration (presenting the global rationale and the performance level) need to be integrated into CAD training procedures. The importance of CAD training is being acknowledged by the radiology community, as evidenced by the fact that CAD training has recently been added to the mammography training programmes of the Radiological Society of North America and the American College of Radiology [79]. However, there is currently no research available that investigates the effects of CAD training on the radiologist-CAD interaction. It would be interesting to evaluate the effects of different training contents on radiologists' trust in CAD. This would contribute greatly to a definition of the optimal CAD training procedure.

Several studies indicate that CAD is more effective for novice than for expert radiologists (e.g. [26,80–82]). Although this is likely due to a ceiling effect (experts are already close to maximum performance, whereas novices have much more room for improvement), it is also possible that there is a fundamental difference in radiologist-CAD interaction between novices and experts. It seems natural to assume that experts have more trust in their own abilities than novices. This could cause them to place relatively little trust in CAD, resulting in a smaller effect of CAD on the performance of the radiologist-CAD team. It is worth studying whether this difference truly exists and whether changes in CAD's output aimed at facilitating more appropriate trust differentially affect novices and experts.

References

[1] C.A. Beam, P.M. Layde, D.C. Sullivan, Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample, Arch. Intern. Med. 156 (1996) 209–213.

[2] W.A. Berg, C. Campassi, P. Langenberg, M.J. Sexton, Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment, Am. J. Roentgenol. 174 (2000) 1769–1777.

[3] H.L. Kundel, C.F. Nodine, D. Carmody, Visual scanning, pattern recognition and decision-making in pulmonary nodule detection, Invest. Radiol. 13 (1978) 175–181.

[4] J.H. Austin, B.M. Romney, L.S. Goldsmith, Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect, Radiology. 182 (1992) 115–122.

[5] A. Pinto, L. Brunese, Spectrum of diagnostic errors in radiology, World J. Radiol. 2 (2010) 377–383.

[6] R. Fitzgerald, Error in radiology, Clin. Radiol. 56 (2001) 938–946.

[7] P.J.A. Robinson, Radiology's Achilles' heel: error and variation in the interpretation of the Röntgen image, Br. J. Radiol. 70 (1997) 1085–1098.

[8] D.L. Renfrew, E.A. Franken, K.S. Berbaum, F.H. Weigelt, M.M. Abu-Yousef, Error in radiology: classification and lessons in 182 cases presented at a problem case conference, Radiology. 183 (1992) 145–50.

[9] J.E. Martin, M. Moskowitz, J.R. Milbrath, Breast cancer missed by mammography, Am. J. Roentgenol. 132 (1979) 737–739.

[10] H. Chan, K. Doi, C.J. Vyborny, R.A. Schmidt, C.E. Metz, K.L. Lam, et al., Improvement in radiologists' detection of clustered microcalcifications on mammograms: the potential of computer-aided diagnosis, Invest. Radiol. 25 (1990) 1102–1110.

[11] S. Nawano, K. Murakami, N. Moriyama, H. Kobatake, H. Takeo, K. Shimura, Computer-aided diagnosis in full digital mammography, Invest. Radiol. 34 (1999) 310–316.

[12] K. Awai, K. Murao, A. Ozawa, M. Komi, H. Hayakawa, S. Hori, et al., Pulmonary nodules at chest CT: effect of computer-aided diagnosis on radiologists' detection performance, Radiology. 230 (2004) 347–352.

[13] F. Li, M. Aoyama, J. Shiraishi, H. Abe, Q. Li, K. Suzuki, et al., Radiologists' performance for differentiating benign from malignant lung nodules on high-resolution CT using computer-estimated likelihood of malignancy, Am. J. Roentgenol. 183 (2004) 1209–1215.

[14] S. Kasai, F. Li, J. Shiraishi, K. Doi, Usefulness of computer-aided diagnosis schemes for vertebral fractures and lung nodules on chest radiographs., AJR. Am. J. Roentgenol. 191 (2008) 260–265.

[15] C.S. White, R. Pugatch, T. Koonce, S.W. Rust, E. Dharaiya, Lung nodule CAD software as a second reader: a multicenter study, Acad. Radiol. 15 (2008) 326–333.

[16] T. Drew, C. Cunningham, J.M. Wolfe, When and why might a computer-aided detection (CAD) system interfere with visual search? An eye-tracking study., Acad. Radiol. 19 (2012) 1260–1267.

[17] R.M. Nishikawa, R.A. Schmidt, M.N. Linver, A. V. Edwards, J. Papaioannou, M.A. Stull, Clinically missed cancer: how effectively can radiologists use computer-aided detection?, (2012).

[18] R.F. Brem, J.M. Schoonjans, Radiologist detection of microcalcifications with and without computer-aided detection: a comparative study, Clin. Radiol. 56 (2001) 150–154.

[19] P.M. Taylor, J. Champness, R.M. Given-Wilson, H.W.W. Potts, K. Johnston, An evaluation of the impact of computer-based prompts on screen readers' interpretation of mammograms, Br. J. Radiol. 77 (2004) 21–27.

[20] P. Taylor, R. Given-Wilson, J. Champness, H.W.W. Potts, K. Johnston, Assessing the impact of CAD on the sensitivity and specificity of film readers., Clin. Radiol. 59 (2004) 1099–1105.

[21] B. Zheng, R.G. Swensson, S. Golla, C.M. Hakim, R. Shah, L. Wallace, et al., Detection and classification performance levels of mammographic masses under different computer-aided detection cueing environments, Acad. Radiol. 11 (2004) 398–406.

[22] D. Gur, J.H. Sumkin, H.E. Rockette, M. Ganott, C. Hakim, L. Hardesty, et al., Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system, J. Natl. Cancer Inst. 96 (2004) 185–190.

[23] B. De Hoop, D.W. De Boo, H.A. Gietema, F. Van Hoorn, B. Mearadji, L. Schijf, et al., Computer-aided detection of lung cancer on chest radiographs: effect on observer performance, Radiology. 257 (2010) 532–540.

[24] M.S. Brown, J.G. Goldin, S. Rogers, H.J. Kim, R.D. Suh, M.F. McNitt-Gray, et al., Computer-aided lung nodule detection in CT: results of large-scale observer test, Acad. Radiol. 12 (2005) 681–686.

[25] N. Petrick, M. Haider, R.M. Summers, S.C. Yeshwant, L. Brown, E.M. Iuliano, et al., CT colonography with computer-aided detection as a second reader: observer performance study, Radiology. 246 (2008) 148–156.

[26] A.A. Povyakalo, E. Alberdi, L. Strigini, P. Ayton, How to discriminate between computer-aided and computer-hindered decisions: a case study in mammography, Med. Decis. Mak. 33 (2013) 98–107.

[27] J.D. Lee, N. Moray, Trust, self-confidence, and operators' adaptation to automation, Int. J. Hum. Comput. Stud. 40 (1994) 153–184.

[28] J. Lee, N. Moray, Trust, control strategies and allocation of function in human-machine systems, Ergonomics. 35 (1992) 1243–1270.

[29] B.M. Muir, Trust in automation: part I. Theoretical issues in the study of trust and human intervention in automated systems, Ergonomics. 37 (1994) 1905–1922.

[30] B.M. Muir, Trust between humans and machines, and the design of decision aids, Int. J. Man. Mach. Stud. 27 (1987) 527–539.

[31] C.D. Wickens, J.G. Hollands, Engineering psychology and human performance, 3rd ed., Prentice Hall, Upper Saddle River, NJ, 1999.

[32] J.D. Lee, K.A. See, Trust in automation: designing for appropriate reliance, Hum. Factors. 46 (2004) 50–80.

[33] M.T. Dzindolet, S.A. Peterson, R.A. Pomranky, L.G. Pierce, H.P. Beck, The role of trust in automation reliance, Int. J. Hum. Comput. Stud. 58 (2003) 697–718.

[34] R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse, Hum. Factors. 39 (1997) 230–253.

[35] M.T. Dzindolet, L.G. Pierce, H.P. Beck, L.A. Dawe, The perceived utility of human and automated aids in a visual detection task, Hum. Factors. 44 (2002) 79–94.

[36] H.P. Beck, M.T. Dzindolet, L.G. Pierce, Automation usage decisions: controlling intent and appraisal errors in a target detection task, Hum. Factors. 49 (2007) 429–437.

[37] L. Wang, G.A. Jamieson, J.G. Hollands, Trust and reliance on an automated combat identification system, Hum. Factors. 51 (2009) 281–291.

[38] R.M. Nishikawa, A. Edwards, R.A. Schmidt, J. Papaioannou, M.N. Linver, Can radiologists recognize that a computer has identified cancers that they have overlooked?, in: Proc. SPIE 6146, SPIE, San Diego, CA, 2006: pp. 1–8.

[39] D.W. De Boo, M. Uffmann, M. Weber, S. Bipat, E.F. Boorsma, M.J. Scheerder, et al., Computer-aided detection of small pulmonary nodules in chest radiographs: an observer study, Acad. Radiol. 18 (2011) 1507–1514.

[40] K.H. Lee, J.M. Goo, C.M. Park, H.J. Lee, K.N. Jin, Computer-aided detection of malignant lung nodules on chest radiographs: effect on observers' performance, Korean J. Radiol. 13 (2012) 564–571.

[41] S. Halligan, D.G. Altman, S. Mallett, S.A. Taylor, D. Burling, M. Roddie, et al., Computed tomographic colonography: assessment of radiologist performance with and without computer-aided detection, Gastroenterology. 131 (2006) 1690–1699.

[42] S.A. Taylor, C. Robinson, D. Boone, L. Honeyfield, S. Halligan, Polyp characteristics correctly annotated by computer-aided detection software but ignored by reporting radiologists during CT colonography, Radiology. 253 (2009) 715–723.

[43] J. Shiraishi, H. Abe, R. Engelmann, M. Aoyama, H. MacMahon, K. Doi, Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance--initial experience., Radiology. 227 (2003) 469–474.

[44] L.J. Skitka, K.L. Mosier, M. Burdick, Does automation bias decision-making?, Int. J. Hum. Comput. Stud. 51 (1999) 991–1006.

[45] M. Itoh, Toward over-trust-free advanced driver assistance systems, Cogn. Technol. Work. 14 (2012) 51–60.

[46] E. Alberdi, A. Povyakalo, L. Strigini, P. Ayton, Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography, Acad. Radiol. 11 (2004) 909–918.

[47] A.A. Povyakalo, E. Alberdi, L. Strigini, P. Ayton, Evaluating "human + advisory computer" systems: a case study, in: Proc. 18th Br. HCI Gr. Annu. Conf., British HCI Group, Leeds, UK, 2004: pp. 93–96.

[48] B. Zheng, M.A. Ganott, C.A. Britton, C.M. Hakim, L.A. Hardesty, T.S. Chang, et al., Breast imaging soft-copy mammographic readings with different computer-assisted detection cuing environments: preliminary findings, Radiology. 221 (2001) 633–640.

[49] F. Beyer, L. Zierott, E.M. Fallenberg, K.U. Juergens, J. Stoeckel, W. Heindel, et al., Comparison of sensitivity and reading time for the use of computer-aided detection (CAD) of pulmonary nodules at MDCT as concurrent or second reader, Eur. Radiol. 17 (2007) 2941–2947.

[50] F.J. Gilbert, S.M. Astley, M.A. McGee, M.G.C. Gillan, C.R.M. Boggis, P.M. Griffiths, et al., Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom national breast screening program, Radiology. 241 (2006) 47–53.

[51] J. Shiraishi, H. Abe, F. Li, R. Engelmann, H. MacMahon, K. Doi, Computer-aided diagnosis for the detection and classification of lung cancers on chest radiographs: ROC analysis of radiologists' performance, Acad. Radiol. 13 (2006) 995–1003.

[52] M. Samulski, R. Hupse, C. Boetes, R.D.M. Mus, G.J. Den Heeten, N. Karssemeijer, Using computer-aided detection in mammography as a decision support, Eur. Radiol. 20 (2010) 2323–2330.

[53] R. Hupse, M. Samulski, M. Lobbes, A. Den Heeten, M.W. Imhof-Tas, D. Beijerinck, et al., Standalone computer-aided detection compared to radiologists' performance for the detection of mammographic masses, Eur. Radiol. 23 (2013) 93–100.

[54] H. Chan, B. Sahiner, M.A. Helvie, N. Petrick, M.A. Roubidoux, T.E. Wilson, et al., Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: an ROC study, Radiology. 212 (1999) 817–827.

[55] L. Hadjiiski, H. Chan, B. Sahiner, M.A. Helvie, M.A. Roubidoux, C. Blane, et al., Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study, Radiology. 233 (2004) 255–265.

[56] K. Horsch, M.L. Giger, C.J. Vyborny, L. Lan, E.B. Mendelson, R.E. Hendrick, Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set, Radiology. 240 (2006) 357–368.

[57] Y. Jiang, R.M. Nishikawa, R.A. Schmidt, C.E. Metz, M.L. Giger, K. Doi, Improving breast cancer diagnosis with computer-aided diagnosis, Acad. Radiol. 6 (1999) 22–33.

[58] Y. Kashikura, R. Nakayama, A. Hizukuri, A. Noro, Y. Nohara, T. Nakamura, et al., Improved differential diagnosis of breast masses on ultrasonographic images with a computer-aided diagnosis scheme for determining histological classifications, Acad. Radiol. 20 (2013) 471–477.

[59] K. Awai, K. Murao, A. Ozawa, Y. Nakayama, T. Nakaura, D. Liu, et al., Pulmonary nodules: estimation of malignancy at thin-section helical CT—effect of computer-aided diagnosis on performance of radiologists, Radiology. 239 (2006) 276–284.

[60] Y. Matsuki, K. Nakamura, H. Watanabe, T. Aoki, H. Nakata, S. Katsuragawa, et al., Usefulness of an artificial neural network for differentiating benign from malignant pulmonary nodules on high-resolution CT, Am. J. Roentgenol. 178 (2002) 657–663.

[61] A.D. Andre, H.A. Cutler, Displaying uncertainty in advanced navigation systems, in: Proc. 42nd Annu. Meet. Hum. Factors Ergon. Soc., Human Factors and Ergonomics Society, Santa Monica, CA, 1998: pp. 31–35.

[62] J.M. McGuirl, N.B. Sarter, Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information, Hum. Factors. 48 (2006) 656–665.

[63] J. Beller, M. Heesen, M. Vollrath, Improving the driver-automation interaction: an approach using automation uncertainty, Hum. Factors. 55 (2013) 1130–1141.

[64] P. Madhavan, D.A. Wiegmann, F.C. Lacson, Automation failures on tasks easily performed by operators undermine trust in automated aids, Hum. Factors. 48 (2006) 241–256.

[65] M. Hartswood, R. Procter, M. Rouncefield, R. Slack, J. Soutter, A. Voss, "Repairing" the machine: a case study of the evaluation of computer-aided detection tools in breast screening, in: Proc. 8th Eur. Conf. Comput. Coop. Work, Kluwer Academic Publishers, Helsinki, Finland, 2003: pp. 375–394.

[66] A. Malich, D. Sauner, C. Marx, M. Facius, T. Boehm, S.O. Pfleiderer, et al., Influence of breast lesion size and histologic findings on tumor detection rate of a computer-aided detection system, Radiology. 228 (2003) 851–856.

[67] J.E. Bahner, A. Hüper, D. Manzey, Misuse of automated decision aids: complacency, automation bias and the impact of training experience, Int. J. Hum. Comput. Stud. 66 (2008) 688–699.

[68] E.H. Shortliffe, R. Davis, S.G. Axline, B.G. Buchanan, C.C. Green, S.N. Cohen, Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system, Comput. Biomed. Res. 8 (1975) 303–320.

[69] L.R. Ye, P.E. Johnson, The impact of explanation facilities on user acceptance of expert systems advice, MIS Q. 19 (1995) 157–172.

[70] M.M. Eining, D.R. Jones, J.K. Loebbecke, Reliance on decision aids: an examination of auditors' assessment of management fraud, Audit. A J. Pract. Theory. 16 (1997) 1–19.

[71] J.L. Herlocker, J.A. Konstan, J. Riedl, Explaining collaborative filtering recommendations, in: Proc. 2000 ACM Conf. Comput. Support. Coop. Work, ACM Press, Philadelphia, PA, 2000: pp. 241–250.

[72] H. Cramer, V. Evers, S. Ramlal, M. Van Someren, L. Rutledge, N. Stash, et al., The effects of transparency on trust in and acceptance of a content-based art recommender, User Model. User-Adapt. Interact. 18 (2008) 455–496.

[73] W. Wang, I. Benbasat, Recommendation agents for electronic commerce: effects of explanation facilities on trusting beliefs, J. Manag. Inf. Syst. 23 (2007) 217–246.

[74] Hologic, ImageChecker CAD [software], (2007).

[75] J. Gurung, A. Maataoui, M. Khan, A. Wetter, M. Harth, V. Jacobi, et al., Automated detection of lung nodules in multidetector CT: influence of different reconstruction protocols on performance of a software prototype, Fortschr Röntgenstr. 178 (2006) 71–77.

[76] K. Marten, A. Grillhösl, T. Seyfarth, S. Obenauer, E.J. Rummeny, C. Engelke, Computer-assisted detection of pulmonary nodules: evaluation of diagnostic performance using an expert knowledge-based detection system with variable reconstruction slice thickness settings, Eur. Radiol. 15 (2005) 203–212.

[77] N. Petrick, B. Sahiner, S.G. Armato, A. Bert, L. Correale, S. Delsanto, et al., Evaluation of computer-aided detection and diagnosis systems, Med. Phys. 40 (2013) 087001.

[78] M.C. Roy, O. Dewit, B.A. Aubert, The impact of interface usability on trust in web retailers, Internet Res. 11 (2001) 388–398.

[79] Z. Huo, R.M. Summers, S. Paquerault, J. Lo, J. Hoffmeister, S.G. Armato, et al., Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use, Med. Phys. 40 (2013) 077001.

[80] Y. Wang, S. Jiang, H. Wang, Y.H. Guo, B. Liu, Y. Hou, et al., CAD algorithms for solid breast masses discrimination: evaluation of the accuracy and interobserver variability, Ultrasound Med. Biol. 36 (2010) 1273–1281.

[81] C. Balleyguier, K. Kinkel, J. Fermanian, S. Malan, G. Djen, P. Taourel, et al., Computer-aided detection (CAD) in mammography: does it help the junior or the senior radiologist?, Eur. J. Radiol. 54 (2005) 90–96.

[82] T. Kobayashi, X.W. Xu, H. MacMahon, C.E. Metz, K. Doi, Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs, Radiology. 199 (1996) 843–848.